An in-silico framework for comparing and validating transcripts predicted from paired-end reads

Anna Paola Carrieri[1,2,*], Stefano Beretta[1], Gianluca Della Vedova[1], Ernesto Picardi[3], Yuri Pirola[1], Raffaella Rizzi[1], Graziano Pesole[2], Paola Bonizzoni[1]

[1] Università degli Studi di Milano-Bicocca, Dip. di Informatica, Sistemistica e Comunicazione (DISCo), Milano (Italy)

[2] Istituto di Biomembrane e Bioenergetica (IBBE)-CNR, Bari (Italy)

[3] Università degli Studi di Bari, Dip. di Bioscienze, Biotecnologie e Scienze Farmacologiche, Bari (Italy)

*corresponding author: a.carrieri3@campus.unimib.it

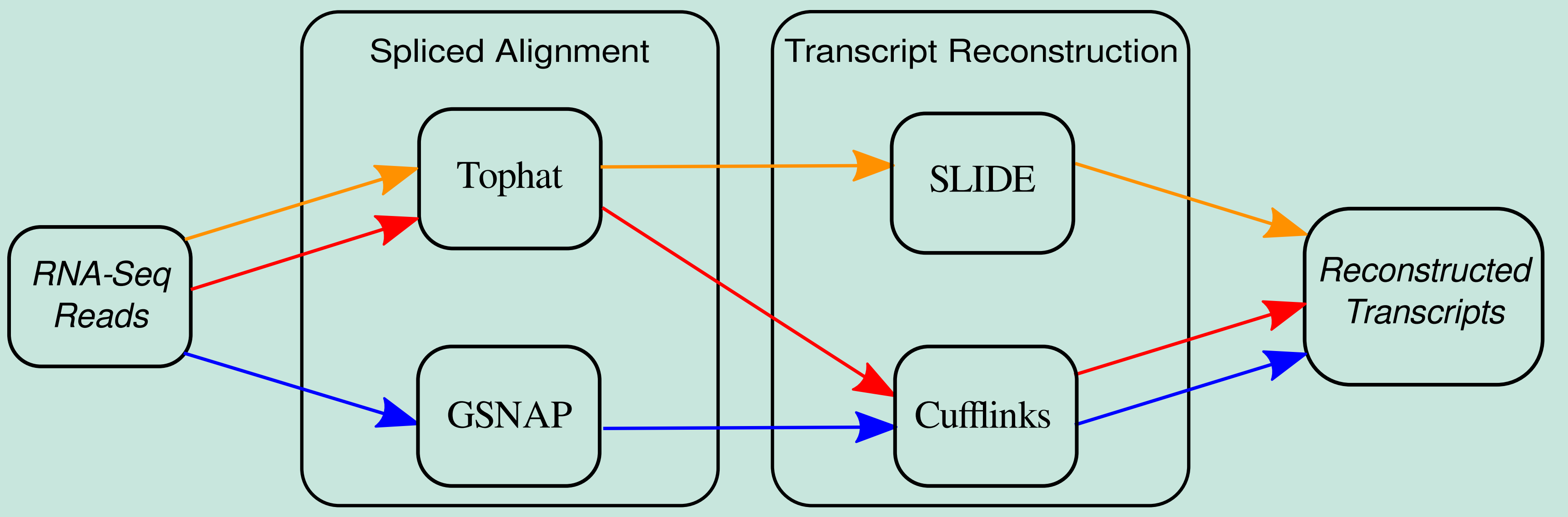
Motivations and Goal

With the advent of high-throughput sequencing of transcriptome (RNA-Seq), different computational methods that use RNA-Seq data to assemble full-length mRNA isoforms have been proposed, albeit not solving completely the problem. We have analyzed some of the most used available tools, evaluating their performance and accuracy.

Methods

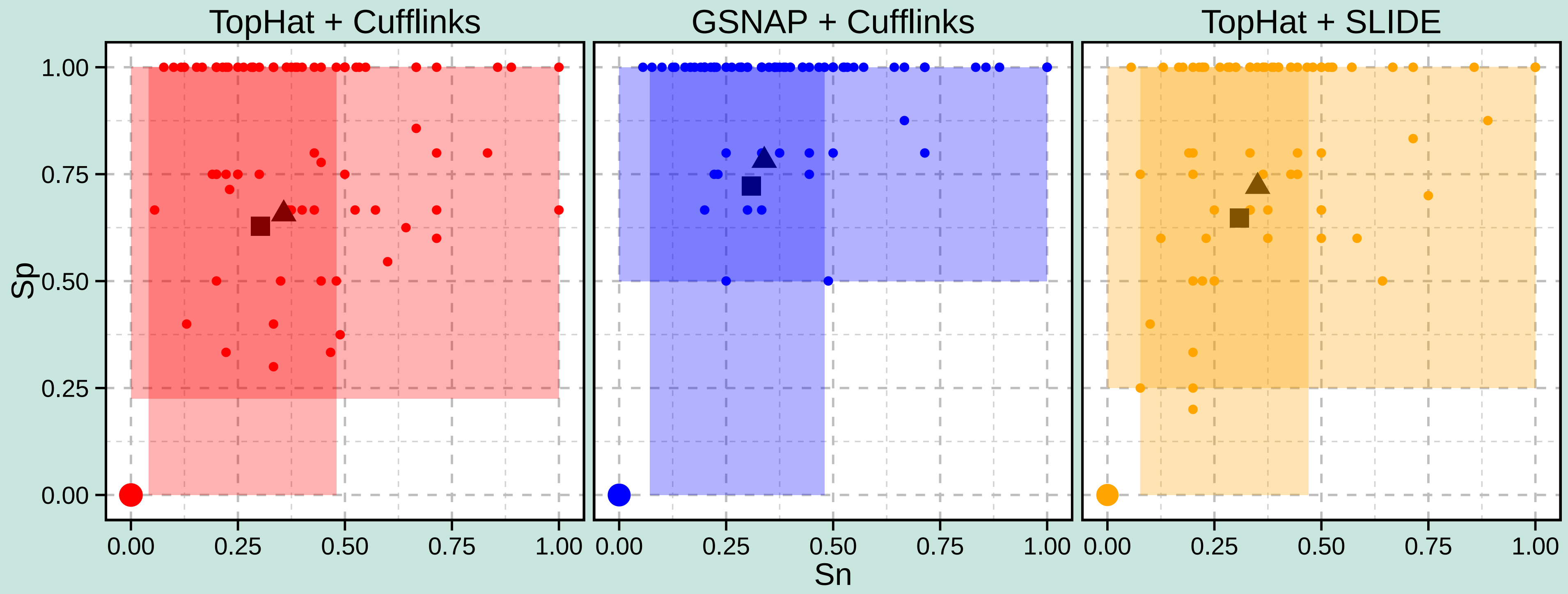
We have developed a program for the automatic execution of several pipelines combining different tools dedicated to the main steps of transcript analysis. More specifically, for splice sites alignment and detection, we considered *Tophat* and *GSNAP*. Afterwards, to assemble the transcripts, the spliced alignments obtained are processed by *Cufflinks* and *SLIDE*, providing all possible pipelines, except for GSNAP+SLIDE which could not be successfully run.

Pipelines



Experiments and Results

We have simulated paired-end reads (2x50bp long at 10x coverage) with ART over annotated isoforms for a set of 112 genes extracted from the 13 ENCODE regions used as training set in the EGASP competition. We have tested all pipelines on those simulated reads in two different situations: one after clustering the reads originating from the same gene and one without clustering (and without any indication of the gene they were originating from.) We used EVAL to analyze the transcript-level specificity and sensitivity of the predictions. The three graphs (one for pipeline) represent the sensitivity and specificity obtained for each gene (point) and for the complete dataset (triangle). The square is the average Sn and Sp over all genes. The table compares each pipeline to determine if one is better than another. The p-values supporting the hypothesis that the first pipeline has a better Sn/Sp than the second is shown.



Hypothesis	p-values Sn	Sp
T+C > G+C	0.875	0.999
T+C > T+S	*0.073*	0.603
G+C > T+C	0.5	**1e-07**
G+C > T+S	*0.055*	**5e-05**
T+S > G+C	0.989	0.999
T+S > T+C	0.981	0.5

Conclusions

Our experimental analysis reveals that using GSNAP instead of TopHat gives more specific predictions with also a minor (but statistically inconclusive) improvement in sensitivity. On the other hand, the comparison of the two pipelines including TopHat has not shown a statistical significant advantage for any pipeline in specificity, but the TopHat+Cufflinks pipeline obtains results that are more sensitive than those computed by TopHat+SLIDE (p-value 0.073). We have been unable to run the complete GSNAP+SLIDE pipeline in our analysis.

We plan to extend our study (i) by introducing some alternatives to EVAL for comparing predictions, (ii) by considering different kinds of simulated data (more coverage levels and/or errors), as well as real data, and (iii) by analyzing in more detail the structure of predicted transcripts since a preliminary study in this direction reveals that the actual methods have various shortcomings in assembling transcripts.

References

[1] C. Trapnell, et al.: *Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation*. Nature Biotechnology 28(5), 516-520 (2010).

[2] J.J. Li, et al.: *Sparse linear modeling of next-generation mRNA sequencing (RNA-Seq) data for isoform discovery and abundance estimation*. Proc. National Academy of Sciences 108(50), 19867-19872 (2011).

[3] S. Beretta, P. Bonizzoni, G. Della Vedova, R. Rizzi: *Reconstructing Isoform Graphs from RNA-Seq data*. IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2012), to appear.

[4] C. Trapnell, L. Pachter, S.L. Salzberg: *TopHat: discovering splice junctions with RNA-Seq*. Bioinformatics 25(9), 1105-1111 (2009).

[5] T.D. Wu, S. Nacu: *Fast and SNP-tolerant detection of complex variants and splicing in short reads*: Bioinformatics 26, 873-881 (2010).

[6] R. Guigò, et al.: *EGASP: the human ENCODE Genome Annotation Assessment Project*. Genome Biol. 7, S2.1-31 (2006).

[7] W. Huang, et al.: *ART: a next-generation sequencing read simulator*. Bioinformatics 28(4), 593-594 (2012).

[8] E. Keibler, M.R. Brent: *Eval: A software package for analysis of genome annotations*. BMC Bioinf. 4(1), 50 (2003).